心理科学进展 2019, Vol. 27, No. 1, 171–180
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2019.00171

# 创造力测评中的评分者效应*

韩建涛[1,2,3] 刘文令[1] 庞维国[1]

([1]华东师范大学心理与认知科学学院, 上海 200062) ([2]安徽师范大学教育科学学院, 芜湖 241000)
([3]巢湖学院文学传媒与教育科学学院, 巢湖 238000)

**摘　要**　创造力测评中的评分者效应(rater effects)是指在创造性测评过程中, 由于评分者参与而对测评结果造成的影响。评分者效应本质上源于评分者内在认知加工的不同, 具体体现在其评分结果的差异。本文首先概述了评分者认知的相关研究, 以及评分者、创作者、社会文化因素对测评的影响。其次在评分结果层面梳理了评分者一致性信度的指标及其局限, 以及测验概化理论和多面 Rasch 模型在量化、控制该效应中的应用。最后基于当前研究仍存在的问题, 指出了未来可能的研究方向, 包括深化评分者认知研究、整合不同层面评分者效应的研究, 以及拓展创造力测评方法和技术等。

**关键词**　评分者效应; 创造力; 主观评分; 评分者认知; 评分者一致性

**分类号**　B841; G449

在社会科学领域, 研究者常常需要以人的主观评判分数, 作为衡量个体工作或行为表现的量化指标。例如, 教师评价学生的作文, 管理者评判员工的工作表现, 在很大程度都依赖于主观判断。然而, 由于每位评分者都有自身独特的风格(idiosyncrasies), 人的主观因素一旦卷入测评, 评分过程中就难免出现偏差。所谓评分者效应(rater effects), 即是由评分者之间的差异, 特别是主观因素差异而对测量结果所造成的影响(Wolfe, 2004; Wolfe & McVay, 2012)。由于评分者因素可能对测验的信效度产生影响, 很多学术组织要求研究者提供相应的理论或实证证据, 以说明其评判结果是合理的(AERA, APA, & NCME, 2014)。

创造性想法或产品不仅是新颖的(novel), 还需是适宜的(appropriate) (Hennessey & Amabile, 2010), 有用的(useful) (Plucker, Beghetto & Dow, 2004; Runco & Jaeger, 2012), 或者是有意义的(meaningful) (Beghetto & Kaufman, 2007)。换言之, 对观念和产品创造力水平的评判, 离不开人们的价值判断。事实上, 主观评定目前也是创造力测评领域广为采用的评定形式(贡喆, 刘昌, 沈汪兵, 2016; Long, 2014a)。当评分者参与创造力测评的情况下, 如何有效描绘、控制测评过程中的评分者效应, 自然也成为创造力研究领域的重要课题(Hung, Chen & Chen, 2012; Long & Pang, 2015)。

目前, 针对评分者效应的研究主要从两个层面展开(Wolfe & McVay, 2012)。一是“潜”层, 主要聚焦于测评本身, 即分析评分背后的认知加工, 揭示不同类型(如不同知识经验水平)评分者在认知特点上存在的差异, 以及影响测评的因素等。二是“表”层, 仅关注评分结果, 分析评分者间的一致性, 利用相应的统计指标和模型对其进行量化, 或利用统计控制的方法, 校正评分偏差。鉴于此, 本文旨在以上述两个层面为分析框架, 针对创造力测评中的评分者效应及其相关研究作以梳理, 以期能够为创造力研究者提供某些参考。

## 1　评分者认知

创造力作为一种高级认知形式, 对其评判肯定是一个复杂的认知加工工程。基于当前研究, 可从两个层面对其加以分析, 一是对特定观点(或产品)创造性的感知和辨识(即创造性观念评价认知); 二是对众多观点创造性的对比和评分(即创造力测评)。

收稿日期: 2017-07-28
* 教育部人文社会科学研究青年基金项目(16YJC190008)。
通信作者: 庞维国, E-mail: wgpang@psy.ecnu.edu.cn

### 1.1 创造性观念评价认知

Runco 和 Smith (1992)将创造性观念评价分为个人评价(intrapersonal evaluation)和人际评价(interpersonal evaluation)两种形式，前者指个体对自己生成观点的评价，后者即对他人观点的评价。其中，个人评价可以发生在创造性认知加工的过程之中，也可以是针对自己最终创造性产物的评价。Mumford, Lonergan 和 Scott (2002)将创造性认知加工的过程中的观念评价定义为估计观点的价值、预判其影响，以及对观点进行修正和精炼等一系列复杂认知活动。与观念生成过程以发散思维为主不同，观念评价过程则以聚合思维为主(Cropley, 2006)，且两者在脑机制上也表现出显著的差别(Ellamil, Dobson, Beeman, & Christoff, 2012)。而观念生成和观念评价也被认为是创造性认知加工的两个主要阶段，这在诸多理论模型中都有所体现(Campbell, 1960; Finke, Ward & Smith, 1992; Sowden, Pringle & Gabora, 2015)。但是，由于发生在创造性认知过程中的观念评价，往往是和观念生成过程交替进行的(Finke et al., 1992)，很难将其单独分离出来加以探讨。因此，目前针对个人评价的研究也是以个体对已生成产品进行评价的方式进行(e.g., Runco & Smith, 1992; Silvia, 2008)。

依据 Runco 和 Smith (1992)的分类，创造力测评中的评价显然属于人际评价的范畴。关于人际评价，已有研究则发现，人们倾向于低估他人想法的创造性水平，特别是高原创性的想法(Licuanan, Dailey & Mumford, 2007)，而偏爱容易理解、符合一般社会规范的观点(Blair & Mumford, 2007)。Mueller, Melwani 和 Goncalo (2012)进一步分析认为，由于创造性想法具有不确定性，很可能是这种不确定感使人们对其产生了消极评价。在其研究中，他们首先启动被试对不确定性的高、低容忍度，然后再让被试对一个高创造性想法进行评判，同时测查了外显、内隐创造力态度。结果显示，大学生被试在不确定性容忍度低的条件下，对想法的创造性评价更低，并且消极的内隐创造力态度在其中起到了中介作用。Mueller, Wakslak 和 Krishnan (2014)还发现，人们在低建构水平(认知表征更加具体化)条件下，更加难以识别想法的创造性，从而表现出对高创造性想法的低估。Zhu, Ritter, Müller 和 Dijksterhuis (2017)新近的研究则发现，相对于精细化加工(deliberative processing)，被试在直觉加工(intuitive processing)条件下选出的想法更具创造性。这些研究表明，人们在识别新颖观点的过程中，可能需要相对整体、抽象和直觉性的思维模式，一方面可减少对相对陌生刺激的不确定感，另一方面促进对创造性观点的理解和辨识。

需注意的是，由于上述研究主要关注个体如何识别特定创造性想法，因此往往以特定、少量、已被评定为高创造性的想法作为评价材料，进而探讨创造性评价的偏向和影响因素等问题。但在创造力测评过程中，评分者要面对众多创造性水平不同的想法，因此其评分认知过程势必更加复杂。

### 1.2 创造力测评及评分认知

#### 1.2.1 创造力测评

当前，创造性的测评主要从四个方面展开，即创造性的过程(creative process)、人(the creative person)、产品(creative products)以及环境(creative environments) (Plucker & Makel, 2010)。近些年来，创造性过程和产品测验被应用得更加深入、广泛(贡喆等, 2016)。创造性思维过程测验主要包括发散思维(Divergent Thinking, DT)测验和顿悟类测验。由于顿悟类测验的问题一般都有明确的答案，不存在评分者效应问题。而发散思维测验和创造性产品测验一般是开放式的，答案不确定、不唯一，就有了人参与评分的需要。因此，接下来的论述将主要围绕 DT 测验中的主观计分和针对创造性产品的同感评估技术(Consensual Assessment Technique, CAT)展开。

在 DT 测验中，早期 Guilford 等已开始使用主观评定的方法对被试答案的原创性(originality)进行评分，并提出了其 3 个指标维度，即非常规性(uncommonness)、远距离性(remoteness)和聪明性(cleverness) (Wilson, Guilford & Christensen, 1953)。譬如，为了评估观点原创性的“聪明性”维度，Guilford 等请 3 位评分者在 0~6 上对被试所生成图片标题进行打分。近些年来，DT 测验的主观计分法又得到进一步发展，已被广泛应用(Benedek, Mühlmann, Jauk, & Neubauer, 2013; Silvia, 2011; Silvia, Martin & Nusbaum, 2009; Silvia et al., 2008)。在创造性产品测验领域，Amabile 首先将产品的创造力定义为合适而独立的评判者赞同其具有创造性的程度(Amabile, 1982)，然后提出了 CAT 要求：所有评判者需具有领域相关经验(即专家)；在不给

予特定标准的情况下，评判者独立进行评判并达到某种程度的一致。Amabile 的研究也表明，CAT 应用于拼贴画、短故事、诗等任务，都有较好的评分者一致性，并与创造力以外的其他维度(如技巧、艺术吸引力)相互独立(Amabile, 1983)。在 Amabile 开创性工作的基础上, CAT 也得到进一步发展(Hennessey, 1994; Kaufman, Baer, Cole, & Sexton, 2008)，同样被广泛应用于创造力实证研究(Long, 2014a)。

尽管 CAT 与 DT 的主观评分是两种不同的创造力测评手段，但二者之间也有相似性之处。首先，它们都是测查众多产品或观点的相对创造力水平，强调待评材料之间的相互对比以及评分顺序的随机。其次，都需评分者主动参与，以其内在的标准或对给定标准的个人理解，对产品或观点的创造性进行评判。因此，评分者认知过程具有一定的不可控性，难免会带来评分者效应问题。

#### 1.2.2 评分过程与标准

与对特定的观点进行评价相比，对众多观点的对比、评分肯定更加复杂。鉴于单纯量化研究的局限，有研究者尝试采用定量和定性研究相结合的方式，对评分者的内在认知过程、特点，以及评分标准等问题进行分析(Long, 2014b; Long & Pang, 2015)。

Long 和 Pang (2015)以六年级学生在科学创造力任务中的反应作为评分材料，选取了创造力研究者(具有创造力领域知识的专家)、教师(具有学生相关知识的专家)和大学生(新手)作为评分者，探讨了其评分特征。基于半结构式访谈的质性分析结果显示，评分者的评分大致可分为三个认知加工阶段：(1)准备(preparing)阶段：评分者阅读评分指导语、理解创造力任务，形成自己对创造力的理解，以作为之后评分的标准；(2)评分(scoring)阶段：评分者一般会通览全部或部分待评的答案，以形成总体性认识，进而依据自己的评分标准对答案的创造力水平进行评定；(3)调整(adjusting)阶段：评分者会将前后的评分进行对比，进一步修改最开始的评分(也有部分评分者不会修改)；比如提高其他人没有提及答案的评分(如果以新颖性作为其评分标准的话)。

另外, Long (2014b)基于科学创造力任务材料，还分析了 CAT 的评分标准问题。研究发现，除了新颖性(novelty)和适用性(appropriateness)，评分者还会依据聪明性(cleverness)、慎思性(thoughtfulness)以及有趣性(interestingness)作为评分的标准；并且，不同评分者所依据的标准或标准组合、赋予每个标准的权重、对同样标准的理解，都会有所差别；针对不同的具体评分任务，同一位评分者也可能会改变自己的评分标准以适应于该任务情境。

针对 DT 测验中被试想法(质量)的评分，评分者被要求依据的标准也并不统一。如 Silvia 等人(2008)借鉴了 Guilford 等原创性的三维度指标，但所评却是每个想法的创造性(creativity)。同样是评观点的创造性，Benedek 等人(2013)则要求评分者依据原创(original)且适用的(useful)标准进行评定。另一些研究，则直接让评分者对想法的原创性(originality) (Fink et al., 2015)，或新颖性(novelty) (Diedrich, Benedek, Jauk, & Neubauer, 2015; Gilhooly, Fioratou, Anthony, & Wynn, 2007)进行评定。

概括看来，目前直接针对评分者认知的实证研究并不多，对其认识尚不够系统和深入。不难想象，随着评分者特征、任务类型、评分情境，甚至是创作者特点的变化，测评过程和结果都可能出现差异。正是因为看到这一点，更多研究者从某一角度切入，具体考察影响创造力测评的各种因素。

## 2 影响创造性测评的因素

### 2.1 评分者的知识经验

按照 CAT 的理论假设，选取专家型评分者是有效测评产品创造力的前提(Amabile, 1983)。专家和新手评分者间对比研究结果也表明，以具有一定知识经验的专家作为评分者或许是必要的。譬如，Kaufman 等人(2008)分别选取专家(诗人)和非专家(大学生)作为评分者，评判了 205 首诗的创造性。结果显示，非专家评分者的评分一致性更低，并且与专家的评分仅有非常弱的相关。

但也有研究为新手评分的可靠性提供了证据。Lu 及其同事以具有多年设计经验的从业人员作为专家，以没有从业经验的设计专业大学生和研究生作为非专家，对比了他们对设计类产品的创造性评判。结果显示，无论是依据 CAT，还是依据给定系列标准的产品创造力测量工具(Product Creativity Measurement Instrument, PCMI)，非专家评分者的评分一致性都更高，并且他们在 PCMI 各标准上的评分，对产品创造力的解释量

更大(Lu & Luh, 2012)。Haller, Courvoisier 和 Cropley (2011)的研究也显示，新手的评分一致性更高。

对于上述不一致结果，Baer, Kaufman 和 Riggs (2009)认为评判材料的领域可能是重要的影响因素。Kaufman, Baer, Cropley, Reiter-Palmon 和 Sinnett (2013)也对比了不同经验水平的评分者(新手、准专家、专家)在不同领域(短故事、工业产品)产品上评分的差异。结果发现，在短故事上，准专家和专家间差异不大，但他们对工业产品的评分结果则不太一致。这说明，某些领域可能更需要专家型评分者。Galati (2015)则认为，需要根据任务的复杂性考虑是否需要选择专家作为评分者：对于高复杂性任务，专家是必要的；而对低复杂性任务，专家和新手的评分结果则趋于相同，因此选择非专家作为评分者显得更加经济。研究也显示，对于 DT 测验这种相对简单、领域一般性的任务，新手评分者即可取得不错的评分效果(Benedek et al., 2013; Silvia et al., 2008)。另外，针对相对复杂的创造性产品任务，也有研究表明，通过对新手进行培训可以实现评分信效度的提升(Storme, Myszkowski, Çelik, & Lubart, 2014)。

但需注意的是，专家和新手的评分差异可能不仅仅体现在评分结果的统计指标上，也可能体现在认知特点上(Kozbelt & Serafin, 2009)。因此，关于知识经验对测评影响的研究，还需和评分者认知相结合，作进一步探讨。

### 2.2 评分者的其他特征

除了知识经验，评分者的人格、智力以及自身的创造力等心理特征也可能影响其测评。Tan 等(2015)以儿童创作的乐高积木产品作为评判材料，以不同专业的大学生作为新手评分者，并测查了评分者的大五人格和日常创造力。结果显示，高宜人性和高日常创造力的评分者，其评分标准更为宽松。Benedek 等(2016)以 DT 任务中被试的想法作为测评材料，考察了评分者人格、智力和言语能力等因素对评判准确性的影响。结果显示，人们倾向于低估观点的创造性水平，但更高水平的开放性、智商和言语能力可减少这种消极偏差，进而提升评判的准确性。这表明，高创造性个体或许更有可能发现、识别出创造性想法。亦即富有创造性的人可能具有双重的技能：在生成更多创造性想法的同时，也更善于识别好的想法(Silvia, 2008)。

Zhou, Wang, Song 和 Wu (2017)新近的研究还发现，在对他人想法创造性评判时，高促进定向的个体对高创造性观念评分更高，而高预防定向的个体则对低创造性观点的评分更高。他们分析认为，一个新观点可能是一种“大胆尝试”，也可能是一种“危险”，而不同调节定向的个体对其的感知和偏好可能会有所不同。此外，Forthmann 等人(2017)探讨了评分者认知负荷对评分一致性的影响。该研究结果显示，更复杂的观念(无论观念集还是单个观念)因包含了更多的信息，会增加评分者的认知负荷，进而造成相互间评分更加不一致。该现象在快照评分法(snapshot，即对每个被试的答案集，给一个整体的创造性分数)和 DT 结果任务(consequences tasks，如“人不需要睡觉会导致哪些后果？”)上，表现得尤为突出。

### 2.3 创作者信息

无论是 CAT 还是 DT 测验的主观评分，待评的观点往往都与其创作主体相分离(Amabile, 1982; Silvia et al., 2008)。在心理测量语境下，这样做可避免创作者相关信息对测验结果的干扰，可以在一定程度上增加评分一致性。但在现实情境中，观点与其作者是密切相联的，因此创作者信息是否会对测评产生影响也成为研究者关注的一个重要问题。

为了探讨创作者年龄信息对创造性测评的影响，Hennessey (1994)曾让 3 组大学生评分者评判儿童、成人所创作不同创造性水平的拼贴画。3 组的条件分别是：正确告知组，即提供真实的创作者年龄信息；年龄信息对调组，即将儿童的作品标注为成人所创作，而成人的作品标注为儿童所创作；无年龄信息组，即不告知创作者的年龄信息。研究结果显示，与不呈现年龄信息相比，呈现何种年龄信息都会提升评分者对儿童所创作拼贴画的创造性评分；但对成人作品，不同组的评分并无差异。这表明，评分者在测评过程中会考虑创作者的特点，进而采取不同的评分策略。Han, Long 和 Pang (2017)的进一步研究表明，评分者对低年龄创作者的观点采择(即设身处地站在创作者的立场上评判)可能在其中起着重要作用。

Kaufman, Baer, Agars 和 Loomis (2010)考察了创作者种族和性别信息对创造性测评的影响。结果显示，大学生评分者对白人女性的诗有轻微的偏爱，但整体上，种族和性别信息对测评结果的

影响不大。然而, Lebuda 和 Karwowski (2013)的研究则显示, 在相对缺少关于待评判产品之间比较信息的情况下, 测评可能更容易受到创作者信息的影响。该研究首先在绘画、科学理论、音乐和诗四个领域选取中等创造力水平的作品, 然后将待评作品分别标注不同虚构的创作者姓名(独特男性名、独特女性名、常见男性名、常见女性名以及匿名)。结果显示, 对于诗和音乐作品, 标注独特名字的作品被评分更高; 整体上, 男性的作品比女性的作品被评判为更有创造性, 对于科学理论的评判更是如此。

### 2.4　社会文化及各因素的交互影响

文化作为人类群体活动的深层心理建构, 对创造力测评的影响主要表现为: 不同文化情境下的评分者对创造力的理解、评判标准、赋予不同标准的权重, 以及对创造性产品的接受程度等都会有所差别。譬如, Lan 和 Kaufman (2012)的研究显示, 美国人倾向于重视新颖的价值, 以及打破常规的创造力类型; 而中国人则倾向于欣赏在限制条件下的创造力, 例如对传统观念的再加工。Hong 和 Lee (2015)的研究也表明, 文化会影响新手评分者对新颖建筑设计的创造性评判; 与美国白人相比, 东亚人对新颖建筑的评分和接受程度更低。这与创造力的跨文化研究结果基本一致, 即东方文化可能更强调想法的适宜性和可行性, 而西方文化则更看重其新颖性(Goncalo & Staw, 2006)。

影响创造力测评的因素具有多元性、相互作用性, 因此近期有研究开始探讨多个因素间的交互作用对创造性评判的影响。Cheng (2016)在研究中以乐高积木作品作为创造力测评任务, 设置了强势(告知作品由乐高狂热者完成)和非强势(告知作品由初学者完成)两种评分条件, 同时还测查评分者的大五人格。其研究结果显示, 情绪稳定性和强势与否存在交互作用: 在非强势条件下, 不同情绪稳定性评分者之间的评分没有差别(都相对宽松); 但在强势条件下, 情绪稳定性低的评分者标准更加严格。这表明, 评分者和创作者之间存在交互影响。Zhou 等人(2017)的研究则揭示, 创造性评分受到评分者(不同调节定向)、观点(不同创造性水平)和情境(损失或收益)三者之间的交互影响。在时间进程上, Kozbelt 和 Serafin (2009)发现对创造性作品的评判具有动态性, 即评分者对创作过程中各阶段的评判是动态变化的; 并且, 作品创造性越高, 其变化规律越复杂, 越难以被预测。鉴于影响创造性测评的因素的复杂性, Birney, Beckmann 和 Seah (2016)近期提出了人−任务−情境三维创造力评判框架, 强调在创造性评估过程中, 综合考虑人、任务、情境因素的共同影响。

### 2.5　小结

综上可见, 人们在评判想法或产品创造性的过程中, 的确会受到诸多因素的影响。有鉴于此, 针对不同的创造性测评手段, 研究者都提出了相应的要求, 以尽量避免其他因素对评分的干扰, 从而实现对评分者效应的控制。譬如, 无论是 CAT 还是 DT 测验的主观评分, 评分者仅对想法或产品进行评判, 并不被告知创作者信息(Amabile, 1982; Silvia et al., 2008)。CAT 还要求评判者先总览所有待评产品, 然后再按随机顺序进行评分, 且对不同维度的评分顺序也应是随机的(Amabile, 1983); 在 DT 测验中, 研究者也需将所有待评观点录入电脑, 并将其随机排列, 以排除书写、反应数量和位置等因素的影响, 并向评分者说明评分所依据的标准以及标准间的关系, 以提升评分的内容和构念效度(Silvia et al., 2008)。

但严格的要求也限制了测评方法的应用范围, 提升了其使用的成本。Kaufman, Beghetto 和 Dilley (2016)即认为, 当前的测评方法本质上是为创造力科学研究而设计, 在应用上有极大的局限性。现实情境中的创造性测评肯定更加复杂, 如在创作者、领域和社会环境等因素上都具有特殊性。因此, 基于上述研究, 为了提升现实情境下创造性评价的信效度, 研究者可能需要综合考虑各种相关因素的影响, 而非简单加以排除。

## 3　评分者效应的量化与控制

由于评分者间的变异是创造性主观评分变异的重要来源, 因此作为支撑测验信效度的一部分, 研究者需提供评分者评分稳定、有效的证据。当然, 在这方面需要提供的最为重要、最为常见的指标是评分者一致性信度。

### 3.1　评分者一致性

作为独立的评分专家, 评分者需要依据自己的评定标准或理解, 进行独立的评判。这时, 一致性即评分者所评分数之间的相关程度。在创造性

主观评分中，评分者一般有多名。为了避免两两相关再取平均，有研究者采用组内相关系数(Intraclass Correlation Coefficient, ICC)来表示测量对象变异在测量分数总体变异中所占的比例(e.g., Fink et al., 2015)。计算ICC需选用不同的模型，而Cronbach's α系数即 ICC 各计算模型中的一个特例 (McGraw & Wong, 1996)。因此在创造力的主观测评中，研究者多直接采用 Cronbach's α 系数。

评分者一致性信度可以有效反映评分者所评分数的稳定性，但稳定并不代表准确。以评分者一致性信度来描绘评分者效应依然存在一些局限：(1)各种一致性系数有其适用条件。如 α 系数的使用前提：每位评分者评分对潜变量的载荷一致，即 tau 等价；评分误差间相互独立，即相关为零。当这些条件无法满足时，信度估计即会出现偏差(Silvia, 2011)。(2)该指标只能反映评分者对作品创造性水平高低顺序评定的一致性，并不反映整体评分可能存在的系统偏差。换言之，即使评分者间的评定很一致，也依然不能确定其所评就一定是创造力。(3)评分者一致性信度只能反映来自评分者变异的大小，并不能从整体上分解测量变异的来源，以明确不同因素对测评结果的影响，以及随着这些变量的变化信度值的变化。(4)其仅能作为一个确定的统计指标，但有时测评的结果已成为既定事实，我们可能更加需要一些统计的方法或技巧进行事后的调整和控制。

正因为注意到评分者一致性信度指标存在的诸多不足，研究者近年来开始尝试以新的统计和测量技术分析评分者效应问题。其中，测验的概化理论和多面 Rasch 模型的应用日趋受到重视。

### 3.2 测验概化理论的应用

针对主观评分，概化理论(Generalizability Theory)将可能影响测评结果的因素，都看成测量的侧面(如评分者侧面、任务侧面等)，进而将测量的总变异加以分解。概化理论的 G 研究，可估计测验的概化系数 g 和可靠性系数 φ。概化理论的 D 研究，则可通过调整全域中各侧面的样本量，进而重新估计测量各变异和信度指标，以为决策提供依据(Long & Pang, 2015; Yang, Oosterhof & Xia, 2015)。

Silvia 等(2008)依据概化理论，对比分析了 DT 测验不同计分方法的可靠性。他们具体考察了三个测量侧面：评分者侧面、计分类型侧面和任务侧面(其中，评分者被看作随机面，而任务和计分类型都被作为固定面)，分析了不同评分方法在不同 DT 任务上的评分者一致性。结果表明，基于主观评定的平均数计分法和 TOP2 计分法(仅对被试自行圈选的两个最有创造性答案计平均分)，在非常规用途(unusual uses tasks)和例举任务(instances tasks)上，评分者的误差变异都不大，测验分数的主要变异可由受测试者的变异所解释；可靠性系数的分析表明，在这两类任务上，当评分者为 2 人时，两种系数基本都达到可接受的水平(0.67~0.84)，并且评分者增加到 3 人，可靠性系数还有适当的提升(可提升 0.05~0.08)；但在结果任务上，两种主观计分法的可靠性都较差，且来自评分者的变异也较大。在 Long 和 Pang (2015)基于 CAT 的研究中，他们也将任务作为固定侧面，将评分者作为随机侧面。结果发现，在科学创造力任务上，来自评分者的变异不大，与测量目的有关的变异同样不大，测量分数更多地由误差变异(受测试者与评分者的交互效应，以及随机误差效应)所决定。信度分析结果则表明，其中一个任务甚至需 10 名以上评分者，才能使三类评分群体的评分可靠性都达到可接受水平(≥ 0.7)。

概言之，测验的概化理论不仅仅可以呈现评分一致性指标，还能够使研究者对测量误差有更全面的把握，同时也能为评分者数量的确定提供依据。此外，无论是 ICC 还是 Cronbach's α，本质上都是概化理论的一种模型特例(Yang et al., 2015)。因此，概化理论作为一种更为灵活的框架，可应用于更为复杂测量情境的信度估计问题。

### 3.3 多面 Rasch 模型的应用

Rasch 模型以潜在特质构建被试在具体测试项目上的特征曲线，将所有潜在特质参数与项目参数定义在同一度量系统上，综合考察被试特质水平、项目难度对正确作答概率的影响，从而提升了参数估计的科学性和灵活性(晏子, 2010; Hung et al., 2012)。多面 Rasch 模型(Many-Facet Rasch Model, MFRM)是对 Rasch 双面模型的扩展，即除了被试者和项目两个侧面，还考虑诸如评分者、评分标准等侧面对评分的影响。不同的 MFRM 模型可被用来回答关于评分数据的不同问题。例如，要锚定项目的难度相同，即可不考虑该侧面，从而形成新的模型。因此，MFRM 模型是评价评分质量的有用工具(Linacre, 1994; Wolfe & McVay,

2012)。

在创造力测评领域, Hung 等(2012)以 MFRM 分析评分者效应的研究显示, 评分者虽没表现出宽大/严格、极端/趋中、光环效应、反应定势/随机效应、安全评分倾向、评分不稳定等评分偏差, 但评分者与各评分标准之间存在交互作用, 即评分者在不同的评分标准上宽严有所不同。Primi (2014)以创造性隐喻(如“骆驼是沙漠中的______”)产品为材料, 以 18 名研究生作为评分者, 让他们对每个产品的质量和灵活性进行评分。尽管研究中的评分者接受了细致的培训, 但基于 MFRM 的分析结果仍显示, 评分者的宽严存在个体差异; 并且, 将宽严度调整为一致, 能提高评分者内部一致性信度指标。

不难发现, MFRM 在某种程度上可将统计指标与评分者认知(如各种评分偏差)联系起来, 并可对评分偏差进行事后控制。因此, 其对评分者效应的分析更为细致, 也为深入理解创造性测评过程提供了更多的信息。另外, 由于 MFRM 以同样的“标尺”量化各种参数, 方便了分数间的等值转换。研究者只需利用一些“锚定项目”, 即可将不同人的评分关联起来, 从而使主观评分的应用更加灵活。

## 4 总结与展望

近年来, 创造力评估中的评分者效应尽管日益受到研究者的重视, 但客观看来, 这一研究领域方兴未艾, 仍存在诸多问题, 有待进一步系统、深入的探讨。如下三个方面, 尤其值得研究者关注。

### 4.1 深化评分者认知研究

综观当前有关评分者认知的研究, 不难发现, 首先相关研究尚比较零散。譬如, 创造性感知(creativity perception)、观念评价(idea evaluation)、观念选择(idea selection)和创造性测评(creativity assessment)等主题目前都已被关注, 但还缺少对它们之间关系的理论分析和实证研究。再加之各具体研究所使用的材料、范式又存在很大的差别, 这进一步增加研究结果间比较和整合的难度。因此, 未来有必要进一步厘清相关概念术语之间的区别和联系, 进而构建更加合理、系统的评分者认知研究框架。

其次, 目前还缺乏对评价和评分认知机制的研究。与有明确答案或相对客观标准的评分相比, 创造性的主观评价受更多不可控因素的影响。例如, 评价者采取的评分标准可能会不同(Long, 2014b), 评价的过程也存在个体差异(Long & Pang, 2015)。但当前的研究还更多停留在揭示现象的层面。因此, 为了更好地以人作为创造力测评的工具, 研究者需要对个体观念评价认知机制有更深入的了解。这包括评分者的评价标准、认知加工过程, 以及记忆与决策系统在其中发挥的作用等。此外, 目前有关创造性评价认知神经机制的研究也非常有限, 这与创造性观念生成领域大量的脑机制研究形成明显对比(Ellamil et al., 2012)。而观念评价和观念生成是紧密相联的, 且两种认知加工本身也存在相互的作用(Hao et al., 2016)。因此, 评分者认知和神经机制的研究不仅对理解评分者效应有参考价值, 对揭示创造性认知加工的本质同样意义重大。

### 4.2 整合不同层面评分者效应的研究

目前, 研究者对创造力测评中评分者效应的“表”层和“潜”层都做了大量探讨, 但在两个层面研究的相互整合上还比较匮乏。在“潜”层上, 研究者探讨了众多因素对创造性评分总体偏向或一致性的影响, 但这并不能反映评分特点的全貌(如特定评分者的评分稳定性、评分量程的使用等) (Hung et al., 2012); 在“表”层上, 有研究者尝试运用现代测量技术对评分结果做更为细致的分析, 却很少涉及对测评认知机制的探讨, 而更多是对量化指标的改进和扩展。因此, 未来研究可尝试将两方面的研究加以整合, 在分析“潜”层认知特点的同时丰富现代测量技术的运用, 做更为细致的评分分析, 这或许可以得到更为全面的结果, 从而加深对评分者效应的理解和认识。

目前, 现代测量技术在创造力测评评分者效应中的应用还相对有限。因此, 这种整合取向也可在一定程度上促进新方法和技术的推广。另外, 很多关于影响测评因素的研究, 它们在任务材料、评分者以及评分方法上都存在巨大差别, 特别是针对创造性产品测评的研究更是如此。而现代测量技术的优势即在于分离各种变异(Long & Pang, 2015)、进行事后统计控制(Primi, 2014), 可为不同研究结果的关联和比较提供新的视角。

### 4.3 拓展创造力测评方法和技术

当前有关创造力的测评方法, 本质上是研究者为了科学地研究创造力而设计(Kaufman et al.,

2016)。其核心目标在于，寻找适当的任务材料，区分个体间创造力水平的差异。因此，研究者需要知道哪些人更适宜作为评分者，需要避免评分情境、创作者信息等对测评结果的干扰。但在现实情境中，如组织管理和学校教育情境，也存在着大量创造力测评的现象。显然，现实情境中的测评受到更多因素的影响，并且评判者在其中扮演的角色也更为重要。因此，有必要进一步丰富现实情境中的创造力测评研究，以便为开发适用于实际创造力评估方法提供参考。

近年来，基于计算机“自动化计分”的评分方法，已开始被研究者尝试应用于创造力测评(Harbison & Haarmann, 2014; Beketayev & Runco, 2016)。创造力的主观评分，同样可借鉴类似的技术手段。譬如，将不同研究中被试创造性想法或作品汇集成大型数据库，评分者即可以基于计算机的自动呈现进行评分。采用这种技术，不仅可以提升主观评分的效率、降低使用成本，而且能在某种程度上减轻评分者的认知负荷，减少可能存在的评分者效应问题。此外，评分还可以构成新的“大数据”，以备后续研究的应用或参考。

## 参考文献


贡喆, 刘昌, 沈汪兵. (2016). 有关创造力测量的一些思考. *心理科学进展, 24*(1), 31–45.

晏子. (2010). 心理科学领域内的客观测量——Rasch 模型之特点及发展趋势. *心理科学进展, 18*(8), 1298–1305.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013.

Amabile, T. M. (1983). *The social psychology of creativity.* New York, NY: Springer-Verlag.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing (2014 Edition).* Washington, DC: AERA.

Baer, J., Kaufman, J. C., & Riggs, M. (2009). Brief report: Rater-domain interactions in the consensual assessment technique. *The International Journal of Creativity & Problem Solving, 19*(2), 87–92.

Beghetto, R. A., & Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts, 1*(2), 73–79.

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology, 12*(2), 210–220.

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts, 7*(4), 341–349.

Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., & Neubauer, A. C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity, 21*, 75–84.

Birney, D. P., Beckmann, J. F., & Seah, Y. Z. (2016). More than the eye of the beholder: The interplay of person, task, and situation factors in evaluative judgements of creativity. *Learning and Individual Differences, 51*, 400–408.

Blair, C. S., & Mumford, M. D. (2007). Errors in idea evaluation: Preference for the unoriginal? *The Journal of Creative Behavior, 41*(3), 197–222.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review, 67*(6), 380–400.

Cheng, K. H. C. (2016). Perceived interpersonal dimensions and its effect on rating bias: How neuroticism as a trait matters in rating creative works. *The Journal of Creative Behavior*. February 16, 2017, Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/jocb.156.

Cropley, A. (2006). In praise of convergent thinking, *Creativity Research Journal, 18*(3), 391–404.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 35–40.

Ellamil, M., Dobson, C., Beeman, M., & Christoff, K. (2012). Evaluative and generative modes of thought during the creative process. *NeuroImage, 59*(2), 1783–1794.

Fink, A., Benedek, M., Koschutnig, K., Pirker, E., Berger, E., Meister, S., … & Elisabeth M. W. (2015). Training of verbal creativity modulates brain activity in regions associated with language- and memory-related demands. *Human Brain Mapping, 36*(10), 4104–4115.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications.*. Cambridge, MA: MIT Press.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity, 23*, 129–139.

Galati, F. (2015). Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings

in assessing creativity. *Creativity Research Journal, 27*(1), 24–30.

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology, 98*(4), 611–625.

Goncalo, J. A., & Staw, B. M. (2006). Individualism–collectivism and group creativity. *Organizational Behavior and Human Decision Processes*, *100*(1), 96–109.

Haller, C. S., Courvoisier, D. S., & Cropley, D. H. (2011). Perhaps there is accounting for taste: Evaluating the creativity of products. *Creativity Research Journal, 23*(2), 99–109.

Han, J. T., Long, H. Y., & Pang, W. G. (2017). Putting raters in ratees' shoes: Perspective taking and assessment of creative products. *Creativity Research Journal, 29*(3), 270–281.

Hao, N., Ku, Y. X., Liu, M. G., Hu, Y., Bodner, M., Grabner, R. H., & Fink, A. (2016). Reflection enhances creativity: Beneficial effects of idea evaluation on idea generation. *Brain and Cognition, 103*, 30–37.

Harbison, J. I., & Haarmann, H. (2014). Automated scoring of originality using semantic representations, *Proceedings of the COGSCI*, *36*, 2327–2332.

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal, 7*(2), 193–208.

Hennessey, B. A., & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology, 61*, 569–598.

Hong, S. W., & Lee, J. S. (2015). Nonexpert evaluations on architectural design creativity across cultures. *Creativity Research Journal, 27*(4), 314–321.

Hung, S. P., Chen, P. H., & Chen, H. C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal, 24*(4), 345–357.

Kaufman, J. C., Baer, J., Agars, M. D., & Loomis, D. (2010). Creativity stereotypes and the consensual assessment technique. *Creativity Research Journal, 22*(2), 200–205.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20*(2), 171–178.

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts, 7*(4), 332–340.

Kaufman, J. C., Beghetto, R. A., & Dilley, A. (2016). Understanding creativity in the schools. In Lipnevich, A. A., Preckel, F., & Roberts, R. D. (Eds.), *Psychosocial skills and school systems in the 21st century* (pp. 133–153). Springer.

Kozbelt, A., & Serafin, J. (2009). Dynamic evaluation of high-and low-creativity drawings by artist and nonartist raters. *Creativity Research Journal, 21*(4), 349–360.

Lan, L., & Kaufman, J. C. (2012). American and Chinese similarities and differences in defining and valuing creative products. *The Journal of Creative Behavior, 46*(4), 285–306.

Lebuda, I., & Karwowski, M. (2013). Tell me your name and I'll tell you how creative your work is: Author's name and gender as factors influencing assessment of products' creativity in four different domains. *Creativity Research Journal, 25*(1), 137–142.

Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. *The Journal of Creative Behavior, 41*(1), 1–27.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd Edition). Chicago, IL: MESA.

Long, H. Y. (2014a). An empirical review of research methodologies and methods in creativity studies (2003-2012). *Creativity Research Journal, 26*(4), 427–438.

Long, H. Y. (2014b). More than appropriateness and novelty: Judges' criteria of assessing creative products in science tasks. *Thinking Skills and Creativity, 13*, 183–194.

Long, H. Y., & Pang, W. G. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity, 15*, 13–25.

Lu, C. C., & Luh, D. B. (2012). A comparison of assessment methods and raters in product creativity. *Creativity Research Journal, 24*(4), 331–337.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Mueller, J. S., Melwani, S., & Goncalo, J. A. (2012). The bias against creativity: Why people desire but reject creative ideas. *Psychological Science, 23* (1), 13–17.

Mueller, J. S., Wakslak, C. J., & Krishnan, V. (2014). Construing creativity: The how and why of recognizing creative ideas. *Journal of Experimental Social Psychology, 51*, 81–87.

Mumford, M. D., Lonergan, D. C., & Scott, G. (2002). Evaluating creative ideas: Processes, standards, and context. *Inquiry: Critical Thinking Across the Disciplines, 22*(1), 21–30.

Plucker, J., Beghetto, R. A., & Dow, G. (2004). Why isn't creativity more important to educational psychologists? Potential, pitfalls, and future directions in creativity research. *Educational Psychologist, 39*(2), 83–96.

Plucker, J. A., & Makel, M. C. (2010). Assessment of

creativity. In Kaufman, J. C. & Sternberg, R. J. (Eds.), *The Cambridge handbook of creativity* (pp. 48–73). New York, NY: Cambridge University Press.

Primi, R. (2014). Divergent productions of metaphors: Combining many-facet Rasch measurement and cognitive psychology in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 8*(4), 461–474.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96.

Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences, 13*(3), 295–302.

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts, 2*(3), 139–146.

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity, 6*(1), 24–30.

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity, 4*(2), 79–85.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I.,... Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 68–85.

Sowden, P. T., Pringle, A., & Gabora, L. (2015). The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & Reasoning, 21*(1), 40–60.

Storme, M., Myszkowski, N., Çelik, P., & Lubart, T. (2014). Learning to judge creativity: The underlying mechanisms in creativity training for non-expert judges. *Learning and Individual Differences, 32*(4), 19–25.

Tan, M., Mourgues, C., Hein, S., MacCormick, J., Barbot, B., & Grigorenko, E. (2015). Differences in judgments of creativity: How do academic domain, personality, and self-reported creativity influence novice judges' evaluations of creative productions? *Journal of Intelligence, 3*(3), 73–90.

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin, 50*(5), 362–370.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31–37.

Yang, Y. Y., Oosterhof, A., & Xia, Y. (2015). Reliability of scores on the summative performance assessments. *The Journal of Educational Research, 108*(6), 465–479.

Zhou, J., Wang, X. M., Song, L. J., & Wu, J. (2017). Is it new? Personal and contextual influences on perceptions of novelty and creativity. *Journal of Applied Psychology, 102*(2), 180–202.

Zhu, Y. X, Ritter, S. M., Müller, B. C. N., & Dijksterhuis, A. (2017). Creativity: Intuitive processing outperforms deliberative processing in creative idea selection. *Journal of Experimental Social Psychology, 73*, 180–188.

## Rater effects in creativity assessment

HAN Jiantao[1,2,3]; LIU Wenling[1]; PANG Weiguo[1]

([1] *School of Psychology and Cognitive Science, East China Normal University, Shanghai* 200062, *China*)

([2] *School of Educational Science, Anhui Normal University, Wuhu* 241000, *China*)

([3] *School of Literature Media and Educational Science, Chaohu College, Chaohu* 238000, *China*)

**Abstract:** Rater effects refer to the impact of different raters' idiosyncrasies in their behaviors on the evaluation results in creativity assessment. Rater effects are due to the difference in raters' cognitive process of the evaluation, which are externally reflected in the difference of their scorings. This article first summarizes the studies of rater cognition and other influencing factors on creativity assessment, including characteristics of raters, information of creators and socio-cultural factors. It further examines inter-rater reliability indexes and their limitations, as well as the applications of Generalization Theory and Many-Facet Rasch Model in quantifying and controlling of rater effects. Finally, this paper specifies directions of future research based on the existing limitations, including deepening the investigation on rater cognition in creativity assessment, integrating the studies of rater effects on different levels, and developing new methods and techniques of creativity assessment.

**Key words:** creativity; subjective scoring; rater effects; rater cognition; inter-rater agreement